10-28-2023

From Mistakes to Insights: Counterfactual Explanations for Incorrect Machine Learning Predictions

Amir Asrzad
University of Massachusetts Lowell, Amir_Asrzad@uml.edu

Xiao-Bai Li
University of Massachusetts Lowell, xiaobai_li@uml.edu

From Mistakes to Insights: Counterfactual Explanations for Incorrect Machine Learning Predictions

Completed Research Paper

Amir Asrzad
University of Massachusetts Lowell
Amir_Asrzad@uml.edu

Xiao-Bai Li
University of Massachusetts Lowell
Xiaobai_Li@uml.edu

ABSTRACT

Machine learning (ML) has revolutionized various industries with powerful predictive capabilities. However, the lack of interpretability in these black box models poses challenges in high-stakes domains like finance, healthcare, and criminal justice. Interpretable machine learning (IML) or explainable AI (XAI) aims to address these challenges by developing methods that provide meaningful explanations for human understanding (Molnar, 2023). By enhancing interpretability, we can establish trust, transparency, and accountability in AI systems, ensuring fairness and reliability in their outputs. Counterfactual explanations have gained popularity as an XAI/IML method (Verma, Boonsanong, Hoang, Hines, Dickerson and Shah, 2022). Unlike traditional methods, counterfactual explanations don't directly explain the "why" behind a decision. Instead, they present alternative scenarios, or counterfactuals, illustrating how changes in inputs or features could lead to different outcomes. For instance, if an ML model predicts a loan default, a counterfactual explanation can advise the applicant on the factors that could secure loan approval. Counterfactual explanations are easy to understand, persuasive, and provide actionable insights (Fernández-Loría, Provost, and Han, 2022), leading to increased research attention (Guidotti, 2022).

Fidelity is a crucial criterion for evaluating XAI/IML methods, measuring their ability to approximate black box model predictions accurately. However, existing approaches solely prioritize fidelity and overlook errors. When a black box model misclassifies an instance, interpretable methods, based on fidelity, mistakenly treat the misclassified result as correct and attempt to explain the incorrect outcome. This misinterpretation has significant implications for subsequent actions, and no existing studies have addressed this issue.

In this study, we address the problem of rectifying and explaining incorrect predictions made by AI and ML models. Our focus is on classification problems with two categorical outcomes: beneficial and adverse. Two types of errors exist: misclassifying beneficial as adverse (b2a) and misclassifying adverse as beneficial (a2b). We distinguish between two types of errors, and our research questions are (1) how to explain misclassifications for individuals when a beneficial class is classified as adverse and (2) for organizations when an adverse class is classified as beneficial. We propose a novel and practical method for providing explanations in misclassified cases using a counterfactual explanation approach applicable to any classification model. Our method involves using a black box model to classify instances and fitting a decision tree, called an explanation tree, based on the black box model's classification results. This tree helps identify the best counterfactual examples for explanations, tailored to individual customers and organizational decision-makers and analysts. This work contributes to machine learning and business analytics research in several ways: (1) We investigate the unexplored problem of rectifying and explaining misclassified outcomes made by ML models. (2) We propose a practical method for providing counterfactual explanations in correctly and incorrectly classified cases. (3) We validate our method through an empirical evaluation study using real-world data.

KEYWORDS

Interpretable Machine Learning, Explainable AI, Counterfactual Explanation, Decision Trees, Misclassification